Predicting Oscar Winners

Iain Pardoe

October 7, 2007

Each year, hundreds of millions of people worldwide watch the televised Oscars ceremony. Can one predict which films and which directors, actors and actresses will win the Oscars? Oscars have been presented for outstanding achievement in film every year since 1928, and are generally recognized to be the premier awards of their kind in the motion picture industry.

The Academy of Motion Picture Arts and Sciences (AMPAS) awards them; almost 6000 AMPAS members vote for the nominees and final winners in categories that include directing, acting and writing. Those who take an Oscar home can have a strong likelihood of having exhibited superlative cinematic creativity or achievement[1].

As well as honoring film-makers, Oscars can boost the box-office performance of nominated and winning films. However, although studies into economic factors show that awards can boost movie revenues, there is little overall association between budget and box office variables and the likelihood of winning an Oscar. Oscars seem to be unaffected by how much a movie costs to make, or by how much it makes at the box office. This article does not consider the economic and aesthetic aspects of movies in relation to the Oscars, but focuses purely on the goal of predicting the winners of the four major awards—picture, director, actor in a leading role, actress in a leading role—from those nominated each year.

This might appear to be a purely frivolous activity, suited merely to providing students and movie-loving members of the public with an entertaining example of applied statistics. However, this work may have something more meaningful to say about the merits of actual cinematic performance and the fairness of the Academy's selection and voting process[2].

In terms of data, since the goal is to predict the eventual winner from a list of nominees, any nominee information that is available before the announcement of the winner is potentially useful. This can include other Oscar category nominations, previous nominations and wins, and earlier awards that the movie has won. I use a discrete choice model to provide annual predictions, and then assess predictive accuracy using one-year-ahead, out-of-sample errors. This modelling approach allows prediction of the four major Oscars from 1938 to 2006 (insufficient information accumulated for years before 1938 make predictions unsatisfactory before this date). The final results reveal interesting insights into just how predictable the four major Oscars are, which factors play an important role in the predictions and how these have changed over time. It is also possible to contrast past winners with an exceptionally low estimated probability of winning with past losing nominees with a very high estimated probability of winning.

Data

All data were obtained from The Internet Movie Database (`www.us.imdb.com`). The variables I used are listed in Table 1. They were used to predict the four major Oscar winners from 1938 to

2006, and also provides data ranges for the predicted years' awards. Each variable was included only for the years in which it provided some predictive power.

"Best Picture" and "Best Director" movies are often also nominated several times in other categories, and past records show that the higher the total number of nominations, the greater are the chances of winning. For example, the median number of nominations for winners of the Best Picture and Best Director Oscars since 1928 is nine, whereas the median number of nominations for losing nominees is six.

If your movie is nominated for "Best Picture" and/or "Best Director" it increases your chances in other categories as well. Only three movies have won the Best Picture Oscar without also receiving a Best Director nomination (most recently *Driving Miss Daisy* in 1989) and only two directors have won a Best Director Oscar for a movie that was not nominated for Best Picture (in 1928 and 1929). Thirteen actors and 26 actresses have won Best Actor in a Leading Role and Best Actress in a Leading Role (hereon referred to as Best Actor and Best Actress, respectively) Oscars for roles in movies that were not nominated for Best Picture (most recently Forest Whitaker for *The Last King of Scotland* in 2006 and Reese Witherspoon for *Walk The Line* in 2005).

The Hollywood Foreign Press Association has awarded its Golden Globes every year since 1944. Since Oscars are presented some time after Golden Globes (up to two months later), the award of a Golden Globe often forecasts the winner of the equivalent Oscar. Since 1943, 34 Best Picture Oscar winners had previously won the Golden Globe for Best Picture (Drama); similarly 10 had previously won the Golden Globe for Best Picture (Musical or Comedy). Thirty-five of the Best Director Oscar winners had already won the Golden Globe for Best Director.

The Directors Guild of America (DGA) has been awarding its honors since 1949 and the Producers Guild of America (PGA) has been rewarding the year's most distinguished producing effort since 1989. Fifty-one of the Best Director Oscar winners since 1949 had already won a DGA

Table 1: Explanatory variables used to predict the four major Oscar winners from 1938 to 2006, including data ranges.

Variable	Picture	Director	Lead actor	Lead Actress
Total Oscar nominations	1938–2006	1939–2006	–	–
Director Oscar nomination	1938–2006	–	–	–
Picture Oscar nomination	–	1944–2006	1939–2006	1939–2006
Golden Globe drama	1946–2006	1945–1950	1944–2006	1944–2006
Golden Globe musical/comedy	1956–2006		1965–2006	1952–2006[a]
Guild Award	1951–2006[b]	1951–2006[c]	1995–2006	1996–2006
Previous Oscar nominations[d]	–	1938–2006	1938–2006	–
Previous Oscar wins[d]	–	–	1939–2006	1938–2006
1st front-running movie	1938–2006	1938–2006	1938–2006	1938–2006
2nd front-running movie	1959–2006	1959–2006	1959–2006	1959–2006
3rd front-running movie	1959–2006	1959–2006	1959–2006	1959–2006

[a] Variable dropped between 1961 and 1972 because standard error greatly exceeded estimate.

[b] Directors Guild of America (DGA) for 1951–1988, Producers Guild of America for 1989–2006.

[c] Separate indicators were not included for Best Director awards from both the Golden Globe and DGA from 1951 onwards because of collinearity between the two awards.

[d] Transformed to natural logarithms.

award. Similarly, 31 of the Best Picture Oscar winners from 1949 to 1988 had already won a DGA award and 10 of the Best Picture Oscar winners since 1989 had already won a PGA award. The Screen Actor's Guild (SAG) introduced its own awards, five statuettes known as "The Actor," in 1994. Since 1994, nine winners of the Best Actor Oscar and 10 of the Best Actress Oscar had already won a SAG award.

Nominees for Director and Best Actor seem to have an *increased* chance of winning the more times they have been *nominated* in previous years. However, nominees for Best Actor and Best Actress seem to have a *decreased* chance of winning the more times they have *won* in previous years. For example, 17% of Best Director Oscar nominees with no previous directing nominations have won the Oscar, whereas 24% of Best Director Oscar nominees with one or more previous directing nominations have won. Twenty per cent of Best Actor Oscar nominees with no previous lead actor nominations have won the Oscar, whereas 22% of Best Actor Oscar nominees with one or more previous lead actor nominations have won. Numbers of previous nominations and/or wins were log-transformed in the prediction models because they are highly skewed.

The indicator variable for the first "front-running movie" allows for a nominee's chance of winning an Oscar to be linked to the fortunes of other nominees for the same movie. Each year there are often a handful of movies considered to be the Oscar front-runners, with multiple nominations in the "high-profile" categories (including picture, director, and acting). To identify these front-runners, the Oscar categories were ranked each year based on previous Best Picture Oscar winners. Next, a "nomination score" was calculated for each nominated movie based on these rankings. The indicator variable then identifies the top front-runner as the movie with the highest nomination score, and takes the value of 1 for all nominees associated with this movie. Indicator variables for the second and third front-running movies were derived similarly.

Some additional variables were considered but ultimately not used: the number of previous Best Director Oscar wins, the number of previous Best Actress Oscar nominations, actor and actress ages, movie genre (e.g. drama, comedy), running time, release date, movie critic ratings and other pre-Oscar awards.

Estimation

The goal is to predict the four major Oscar winners for each year from 1938 to 2006 using any information on the nominees that is available before the announcement of the winner. This can be framed as a series of discrete choice problems with one winner selected in each category each year from a discrete set of nominees (five since 1945). In this particular discrete-choice application, the explanatory variables take different values for different response (nominee) choices. McFadden[3] proposed a discrete-choice model for just such a case.

For experiment i and response choice j, let $\boldsymbol{x}_{ij} = (x_{ij1}, \ldots, x_{ijp})^T$ denote the values of p explanatory variables, and let $\boldsymbol{x}_i = (\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{ip})$. Conditional on the choice set C_i for experiment i, the model for the probability of selecting choice j is

$$\Pr(Y = j|\boldsymbol{x}_i) = \frac{\exp(\boldsymbol{\beta}^T \boldsymbol{x}_{ij})}{\sum_{h \in C_i} \exp(\boldsymbol{\beta}^T \boldsymbol{x}_{ih})},$$

where Y is the categorical response variable representing the winning nominee. For each pair of choices a and b, this model has the logit form

$$\log[\Pr(Y = a|\boldsymbol{x}_i)/\Pr(Y = b|\boldsymbol{x}_i)] = \boldsymbol{\beta}^T(\boldsymbol{x}_{ia} - \boldsymbol{x}_{ib}).$$

Conditional on the choice being a or b, a variable's effect depends on the difference in the variable's values for those choices. If the values are the same, then the variable has no effect on the choice between a and b. Thus McFadden originally referred to this model as a conditional logit model, although it is now more commonly called a multinomial logit model.

Such models can be fit with a variety of statistical software packages. For reasons of flexibility, convenience, and familiarity, Bugs was used here for model estimation, with R used to process data and results.

All of the data available before the announcement of the 1938 Oscars were fed into a mathematical model; the output was a prediction of the winners for that year. Then, the actual outcome of the 1938 Oscars was appended to the previous dataset, and used to fit a new model, to be used to predict the winners of the 1939 Oscars. The process repeats, adding new variables as they become available, up to the 2006 Oscars. To assess the predictive accuracy of the analysis, one-year-ahead, out-of-sample errors were used.

Results

Using the modelling approach just described, 190 of the 276 Best Picture, Director, Actor, and Actress Oscar winners from 1938–2006 were correctly identified, corresponding to an overall prediction accuracy of 69%. As more data has become available over recent years, prediction accuracy has improved. For example, the overall prediction accuracy for the last 30 years (1977–2006) is 95 correct predictions out of 120, or 79%. Figure 1 summarizes overall results across the four categories. Overall, the Best Director Oscar has been the most predictable, then the Oscars for Best Picture (until recently), Best Actor, and Best Actress, respectively. Each of the categories has become more predictable over time, particularly Best Actress, which was very hard to predict up

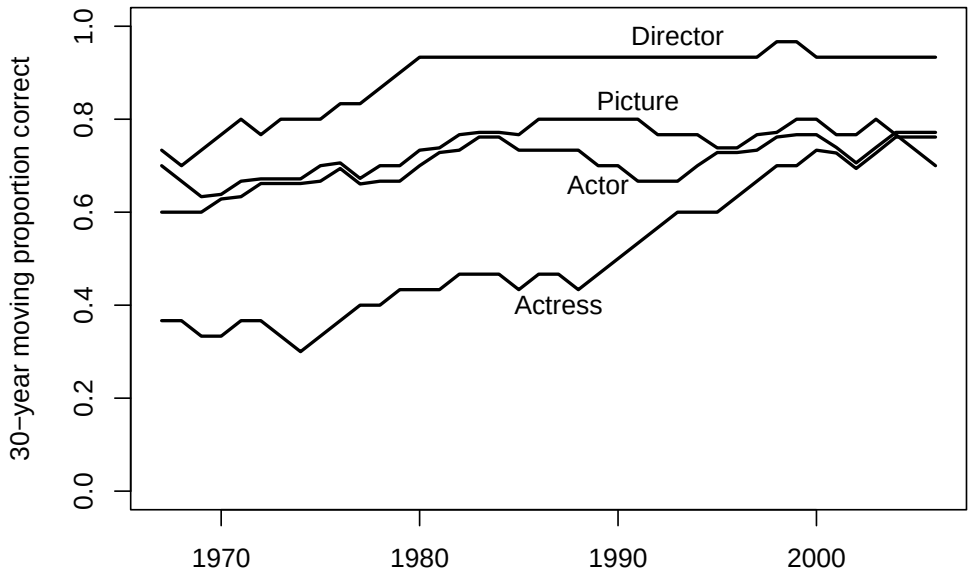

Figure 1: Thirty-year moving averages of the proportion of correct predictions in each of the four major Oscar categories. The moving average values are placed at the ends of the 30-year periods, e.g. at the far right of the graph the proportions of correct predictions over the period 1977–2006 are 93% for Best Director, 77% for Best Actor, 77% for Best Actress, and 70% for Best Picture.

until the early 1970s.

The roles of the explanatory variables in helping to predict Oscar winners have changed over time as is illustrated in Figure 2. The importance of receiving a Best Director nomination (for Best Picture nominees) or a Best Picture nomination (for Best Director, Actor, or Actress nominees)

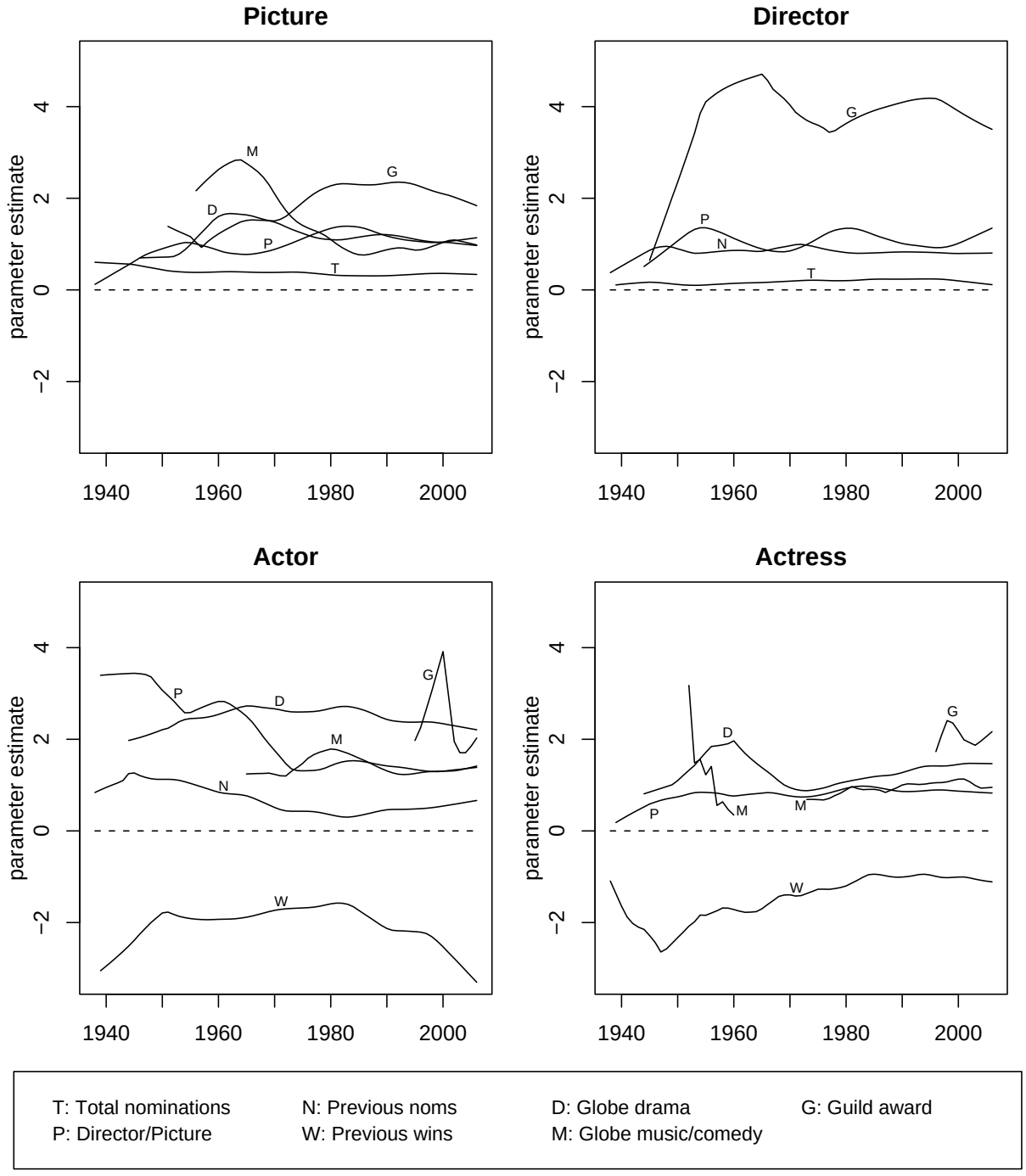

Figure 2: Smoothed parameter estimates for the explanatory variables for each of the four major Oscar categories. The explanatory variables are described in the text.

has tended to increase over time (except perhaps for actors), as shown by the trends in the lines labelled P. Previous nominations have remained approximately equally important for Best Director nominees, but were more important for Best Actor nominees in the past then they have been more recently (lines labelled N). Previous wins seemed to hurt Best Actor nominees less in the 1960s and 1970s than in the 1940s and more recently; however previous wins have tended to become less important for Best Actress nominees over time (lines labelled W).

The Golden Globes have remained useful predictors of future Oscar success since their inception. The changing fortunes of dramas (labelled D) and musicals and comedies (labelled M) can be traced in Figure 2, with musicals and comedies holding an advantage over dramas in the 1960s with respect to Best Picture wins, but with acting wins tending to favor dramas, particularly for males. Guild awards have clearly enabled quite accurate prediction of Best Director winners, and, to a lesser extent, Best Picture winners (lines labelled G). Early indications suggest that SAG awards will be just as helpful in predicting acting wins.

The impact of the total number of Oscar nominations (lines labelled T) on prediction of the Best Picture and Director Oscars remains reasonably steady. Since the total number of nominations has ranged in the past between 1 and 14, this variable is more influential than it appears to be in the graphical illustrations of Figure 2 (which show impacts of the number of nominations increasing by *one*). The impacts of the "front runner" variables—which cut across all four categories—are not shown in Figure 2 (they appeared to be less important, having estimates with smaller magnitudes and larger standard errors).

The analysis also reveals which past nominees have really upset the odds (winners with low estimated probability of winning), and which appear to have been truly robbed (losers with high estimated probability of winning). Table 2 provides details of the three "most surprising" outcomes

Table 2: Three outcomes in each of the major categories with the smallest estimated win probabilities for the actual winner relative to the predicted winner.

Year	Winner	Prob	Predicted	Prob
Best Picture				
1948	*Hamlet*	0.01	*Johnny Belinda*	0.97
2004	*Million Dollar Baby*	0.01	*The Aviator*	0.97
1981	*Chariots of Fire*	0.01	*Reds*	0.88
Best Director				
2000	Steven Soderbergh	0.01	Ang Lee	0.95
1968	Carol Reed	0.02	Anthony Harvey	0.97
1972	Bob Fosse	0.03	Francis Ford Coppola	0.96
Best Actor				
2001	Denzel Washington	0.00	Russell Crowe	0.99
1968	Cliff Robertson	0.00	Peter O'Toole	0.88
1974	Art Carney	0.02	Jack Nicholson	0.87
Best Actress				
2002	Nicole Kidman	0.07	Renée Zellweger	0.90
1985	Geraldine Page	0.07	Whoopi Goldberg	0.70
1950	Judy Holliday	0.09	Gloria Swanson	0.76

in each category (based on the model results). A complete listing of the results is available at `http://lcb1.uoregon.edu/ipardoe/oscars/`—the site is updated in February each year.

Discussion

Discrete choice modelling of past data on Oscar nominees in the four major categories—Best Picture, Director, Actor, and Actress—enables prediction of the winners in these categories with a reasonable degree of success (in recent years: approximately 70% for Best Picture, 93% for Best Director, 77% for Best Actor, and 77% for Best Actress). The analysis could also be extended to other Oscar categories, such as the supporting acting and screen-writing awards.

A limitation of the model is that it can give very extreme predictions that cannot (of course) account for unmeasured factors. I recall the surprise of Denzel Washington winning over Russell Crowe in the 2001 Oscar race for Best Actor, but the surprise was not as extreme as implied by the model predictions in Table 2. Another example is *Brokeback Mountain* failing to win Best Picture for 2005 (after winning a Golden Globe and the PGA award). Again, the surprise of *Crash* winning instead was not as extreme as implied by the model predictions of 0.03 probability for *Crash* versus 0.90 probability for *Brokeback Mountain*—but the model was unable to make use of the late surge that *Crash* made (in unquantifiable "Hollywood buzz" terms) as the Oscars Ceremony drew near.

Further exploration of the results reveals additional findings on the predictability—or lack thereof—of winning an Oscar. For example, there has been much media speculation about legendary individuals who have never won an Oscar, such as Alfred Hitchcock with five directing nominations, Peter O'Toole with eight lead actor nominations, Richard Burton with six lead actor nominations, and Deborah Kerr with six lead actress nominations. Of these, the unluckiest was probably O'Toole who came closest to winning in 1968 (for *The Lion in Winter* in which he took the role of King Henry II; our model gave him an 88% probability of winning) and 1964 (83% probability, for *Becket*, in which, remarkably, he was also portraying Henry II). Kerr came close in 1956 (for *The King and I*, with 72% probability), as did Burton in 1977 (62% probability for his role in *Equus*), while Hitchcock's nearest miss was for *Rebecca* in 1940 (42% probability). Hitchcock, Kerr, and O'Toole were subsequently awarded honorary Oscars.

This article is an updating of previously published work[4]. Finally, as mentioned earlier, Pardoe and Simonton[2] provide deeper insights into the modelling process and potential uses for the analysis.

References

1. Simonton, D. K. (2004). Film awards as indicators of cinematic creativity and achievement: A quantitative comparison of the Oscars and six alternatives. *Creativity Research Journal 16*, 163–172.
2. Pardoe, I. and D. K. Simonton (2008). Applying discrete choice models to predict Academy Award winners. *Journal of the Royal Statistical Society, Series A 171*, 375–394.
3. McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. In P. Zarembka (Ed.), *Frontiers in Econometrics*, pp. 105–142. New York: Academic Press.
4. Pardoe, I. (2005). Just how predictable are the Oscars? *Chance 18*(4), 32–39.